Top-Down System for Multi-Person 3D Absolute Pose Estimation from Monocular Videos

Amal El Kaid, Denis Brazey, Vincent Barra, Karim Baïna

sensors

MDPI

Article

Top-Down System for Multi-Person 3D Absolute Pose Estimation from Monocular Videos

Amal El Kaid [1,2,3,*], Denis Brazey [3], Vincent Barra [1] and Karim Baïna [2]

[1] Université Clermont-Auvergne, CNRS, Mines de Saint-Étienne, Clermont-Auvergne-INP, LIMOS, 63000 Clermont-Ferrand, France; vincent.barra@limos.fr
[2] Alqualsadi Research Team, Rabat IT Center, ENSIAS, Mohammed V University in Rabat, Rabat 10112, Morocco; karim.baina@ensias.um5.ac.ma
[3] Société Prynel, RD974, 21190 Corpeau, France; dbrazey@pryntec.com
* Correspondence: amal.el_kaid@doctorant.uca.fr

Abstract: Two-dimensional (2D) multi-person pose estimation and three-dimensional (3D) root-relative pose estimation from a monocular RGB camera have made significant progress recently. Yet, real-world applications require depth estimations and the ability to determine the distances between people in a scene. Therefore, it is necessary to recover the 3D absolute poses of several people. However, this is still a challenge when using cameras from single points of view. Furthermore, the previously proposed systems typically required a significant amount of resources and memory. To overcome these restrictions, we herein propose a real-time framework for multi-person 3D absolute pose estimation from a monocular camera, which integrates a human detector, a 2D pose estimator, a 3D root-relative pose reconstructor, and a root depth estimator in a top-down manner. The proposed system, called Root-GAST-Net, is based on modified versions of GAST-Net and RootNet networks. The efficiency of the proposed Root-GAST-Net system is demonstrated through quantitative and qualitative evaluations on two benchmark datasets, Human3.6M and MuPoTS-3D. On all evaluated metrics, our experimental results on the MuPoTS-3D dataset outperform the current state-of-the-art by a significant margin, and can run in real-time at 15 fps on the Nvidia GeForce GTX 1080.

Keywords: 3D multi-person pose estimation; absolute poses; camera-centric coordinates; computer vision; artificial intelligence; deep-learning

Citation: El Kaid, A.; Brazey, D.; Barra, V.; Baïna, K. Top-Down System for Multi-Person 3D Absolute Pose Estimation from Monocular Videos. *Sensors* **2022**, *22*, 4109. https://doi.org/10.3390/s22114109

Academic Editors: Tomasz Krzeszowski, Adam Świtoński, Michal Kepski, Carlos Tavares Calafate and Gregorij Kurillo

Received: 15 March 2022
Accepted: 24 May 2022
Published: 28 May 2022

1. Introduction

Human pose estimation (HPE) is a popular task in computer vision. It aims to predict and track the location of joints (e.g., elbow, wrist) or body parts of one or more human bodies; it associates them with segments in graphical form (from an image or sequence of images) to represent the human's orientation and it describe the actual posture. This is an important process for understanding human behavior and human–computer interactions. An example of a human posture skeleton is illustrated in Figure 1.

With human pose estimation, tracking a person or multiple people in real space can be done at an incredibly granular level. This powerful capability unlocks a wide range of industrial applications [1–8], including gaming, animation, motion transfer, augmented reality, human–robot cooperation and training, biomechanical analysis for medical/healthcare, sports fields, gesture control, autonomous driving, human fall detection, action prediction, security and surveillance, etc.

Pose estimation can be performed in two ways: in a two-dimensional space to predict XY image coordinates or in a three-dimensional space to predict the XYZ camera or world coordinates. However, most real-life applications require depth estimation, which provides informative knowledge since 2D poses are often confusing. They can appear identical when in fact they represent completely distinct poses. This makes activity recognition difficult and leads researchers to employ 3D pose estimation.

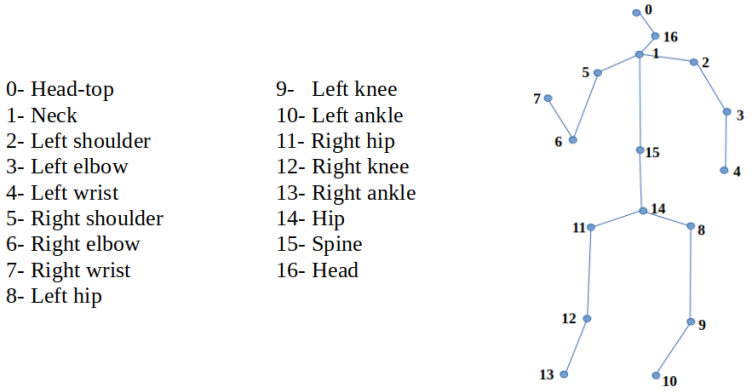

Figure 1. 3D Skeleton model in MuPoTS-3D format and joints names.

Recently, 3D root-relative human pose estimation has shown remarkable progress. Several methods [9–14] propose alleviating the problem by using multi-view images or videos as input. However, multi-view observations are expensive to obtain in daily life scenarios. Thus, the use of 3D human pose estimations from monocular images or videos is in high demand. State-of-the-art approaches that use monocular data [15–22] usually decouple the problem into two main phases: 2D pose estimation for joint detection and localization in the image space, and then lifting of the 2D pixel coordinates to 3D keypoint-position predictions in the camera space. In our research, we followed the same strategy and focused on the second phase, i.e., the 3D pose reconstruction from a sequence of 2D keypoints. Two-dimensional (2D) pose estimation is a popular vision problem that has been studied in many works, e.g., [23–28] and has been greatly improved especially using the deep learning paradigm.

Indeed, 3D pose estimation approaches show promising results on single-person datasets, such as Human3.6M [29] and HumanEva-I [30]. However, they do not perform well in multi-person scenarios, which are the most common cases in real-world applications and surveillance systems. The distances between people can be crucial in the analysis and recognition of their interactions. This introduces the absolute pose [31–33], which aims to locate the root joint (key central point of the person) and estimate its distance from the camera. At present, the 3D multi-person pose estimation still faces a great challenge. When possible, stereo vision calibration is used to determine the exact position of a person from images taken from different points of view. However, these kinds of data are not always available, and they significantly raise the overall costs of the process procedures. Moreover, acquiring such data is impractical in real-time system applications, as we seek to optimize the amount of data that must be captured and processed. This shows the gap between scientific literature and real-world requirements.

The purpose of this study was to present a framework that could accomplish more accurate and robust 3D multi-person pose estimations from a monocular video, from these circumstances and industrial constraints. Thus, we propose an integrated top-down approach that combines GAST-Net for reconstructing 3D root-relative keypoints from 2D keypoints and RootNet for estimating root depth from human bounding boxes. It generates an appropriate 3D multi-skeleton estimation result from a monocular video while maintaining low computational costs and short execution times.

Basically, the system is the result of a series of improvements that boost accuracy by more than 8.8 percentage points on 3D-PCK*abs* on the MuPoTS-3D [34] dataset, when compared to the approaches in the literature [31–33,35,36].

Examples of results from our whole framework are illustrated in Figure 2.

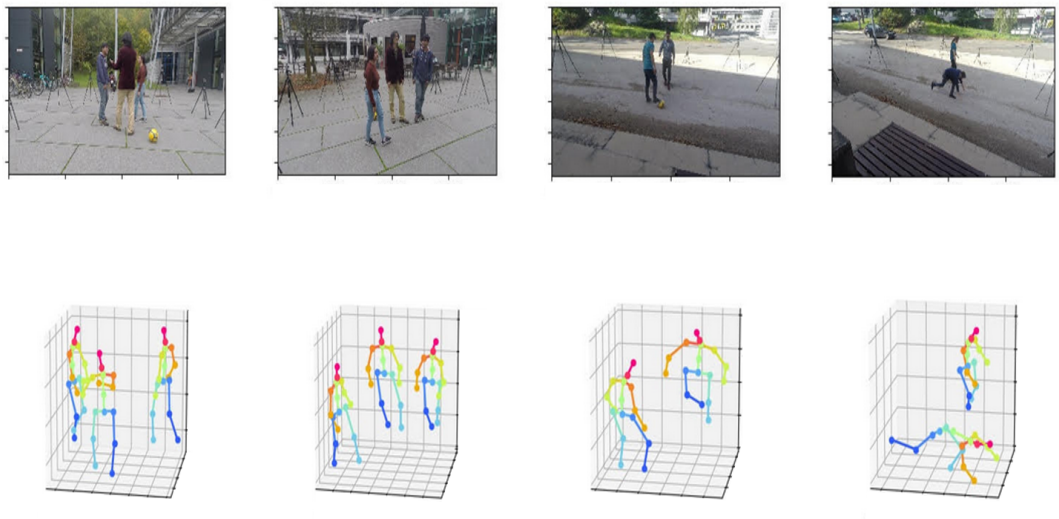

Figure 2. Examples of 3D absolute poses resulting from our whole framework.

The main contributions of this work can be summarized as follows:

- The proposal of an integrated top-down framework based on a modified GAST-Net and RootNet networks for multi-person 3D pose estimation from a monocular RGB video in a short execution time.
- Outperforming existing 3D multi-person absolute pose estimation methods in a MuPoTS-3D dataset by more than 8.8 percentage points on 3D-PCK$_{abs}$ and by more than 12.6 percentage points on AP_{25}^{root} .

The paper is organized as follows. Section 2 illustrates the review of conventional literature on 3D pose estimation based on different levels: the input type (video), the number of instances (multi-person), and the approach following the 3D root-relative pose estimation (two-stage approach). Section 3 demonstrates the proposed framework methodology. Section 4 explains the implementation details, the results and discussion. Section 5 provides a conclusion of the work.

2. Related Works

2.1. Two-Stage Pose Estimation

Several works [22,37–43] apply deep neural networks on 3D pose estimation tasks to learn the direct mapping between RGB images and their corresponding 3D poses in one stage. However, this needs labeled data for supervised training, usually impractical out of MoCap labs. Unsupervised learning algorithms require sophisticated architectures with high computation costs, which are impractical too in realistic applications. To this end, Martinez et al. [44] introduced a two-stage prediction approach. They first predicted the 2D pose from the image and then lifted 2D joint coordinates to the 3D space via a fully connected residual network. Fang et al. [45] introduced a model to encode the mapping function of the human pose from 2D to 3D by explicitly encoding the human body configuration with pose grammar. To improve the generalization of the trained 2D-to-3D pose estimator, Gong et al. [46] proposed a pose augmentation framework (PoseAug) exploiting a differentiable augmentation module based on a neural network. In Ref. [47], the authors created a shape dictionary by collecting all 3D poses in the training set to be aligned by the Procrustes method, to concisely summarize the variability in training data and enable a sparse representation. A convex approach was then proposed to jointly estimate the coefficients of the sparse representation. The same authors [48] predicted the uncertainty heatmaps of the 2D joint locations, then combined these maps with a sparse model of a 3D human pose to retrieve the 3D pose via an EM algorithm. Ref. [49] adopted

a large library of 2D keypoints and their 3D representations to match the depths of the 2D poses estimated by the k-nearest neighbor algorithm. Hossain et al. [50] proposed two 2-layered normalized LSTM networks with residual connections to leverage temporal information for lifting 2D joint locations to 3D positions.

2.2. Video Pose Estimation

Although 3D coordinates can be determined from a single image, temporal algorithms used in videos have better accuracies than simple frame-by-frame approaches. Most works deploy recurrent neural networks (RNNs) [50,51] to exploit temporal information. Long short-term memory networks (LSTMs) [52] are the most widely used RNN architectures for learning long-term dependencies in pose estimation problems because of their ability to preserve information over time. In [51], propagating LSTM networks (p-LSTMs) were proposed to estimate depth information from 2D keypoints. Ref. [53] presented a two-part spatial–temporal convolutional LSTM model (ST-CLSTM) to capture spatial features and temporal consistency between frames. The authors used ST-CLSTM as the generator and a 3D CNN as the discriminator to output the temporal loss from the estimated and ground truth depth sequences. AnimePose [54] used Scene-LSTM to estimate the person's temporal trajectory and track overlapping postures in obscure frames based on their predictions in prior frames. Temporal convolutional networks (TCNs) [55], on the other hand, give additional benefits, such as convolution sharing and low memory requirements for training; this is very advantageous when dealing with extended input sequences. TCN evaluation and training are hence faster than with RNN. As a result, they are becoming increasingly employed in pose estimation [35,37–39,56], especially in real-time systems [57,58]. Moreover, Ref. [39] proposed employing dilated temporal convolutions in a fully convolutional model; moreover, [59] used it as an automatic framework for semantic motion segmentation. Li et al. [60] captured long-range dependencies using transformer-based architecture.

2.3. Spatial–Temporal Graph Convolution Network

Despite the acquired temporal information's ability to anticipate smoother poses, the depths and self-occlusions remain ambiguous. A graph convolutional network (GCN) was used to exploit the spatiotemporal information that allowed to lower these ambiguities. GCNs have greatly improved 3D human pose estimations by representing the human skeleton as an undirected graph. The spatial–temporal graph convolutional network (ST-GCN) [61] was the first approach to use graph CNNs for skeleton-based action recognition. Zhou et al. [22] developed the semantic graph convolutional network (SemGCN) for the 3D human pose regression challenge. The SemGCN aims to learn by capturing semantic information, such as local and global node relationships through end-to-end training. The graph attention spatiotemporal convolutional network (GAST-Net) [57] also combines common convolutional networks to integrate the spatiotemporal information. GAST-Net comprises two types of graph attention blocks: a local spatial attention network (to model the hierarchical and symmetrical structures of the human skeleton) and a global spatial attention network (to extract global semantic information and better encode the human body's spatial characteristics). Cai et al. [62] developed an undirected graph to model the spatial–temporal connections between distinct joints for 3D single-person pose estimation from video data. In Ref. [32], the authors utilized a graphical neural network (GNN) to efficiently aggregate the features corresponding to the different types of articulation, where each type was represented by a graph node. The GCNs based on directed graphs were also adopted by Cheng et al. [35] to model human joint GCNs that refine potentially imperfect poses obtained from 2D pose heatmaps, and human bone GCNs, to model bone connections. The authors also used two TCNs to estimate the 3D root-relative pose and the absolute root depth. Finally, the dynamic graph convolutional module (DGCM) [63] applied GCN for a multi-person 2D pose estimation framework.

2.4. Multi-Person 3D Pose Estimation

Only a few studies were conducted on 3D multi-person pose estimation from a single RGB image. Generally, existing methods can be divided into two categories: top-down and bottom-up approaches.

Top-down 3D human pose estimation methods [64–66] commonly use human detection as an essential part to crop each person in a bounding box and then estimate person-centric 3D full-body joints [31,39,58]. These methods show promising performances, but their main drawbacks still involve the independent detection and process of each person. Hence, they are likely to suffer from inter-person occlusions and close interactions. Rogez et al. [65,67] introduced LCR-Net, which classified bounding boxes generated into a set of K-poses, refined using a regressor. The architecture contains three stages that share the convolutional feature layers and are jointly trained. Likewise, Benzine et al. [68] proposed the pose estimation and detection anchor-based network (PandaNet), an anchor-based single-shot approach. The network predicts the 2D/3D pose regression into a single forward pass for each bounding box detected in a given image.

To predict camera-centric, Moon et al. [31] processed each cropped person's image independently. They produced root-relative 3D joints using PoseNet [21] and estimated the pelvis keypoint localization of each person using the RootNet model. Similarly, hierarchical multi-person ordinal relations (HMOR) [69] is a coarse-to-fine architecture that hierarchically estimates multi-person ordinal relations through instance-level, part-level, and joint-level. The end-to-end HDNet architecture [32] follows the same pipeline, extract pose, and depth data using a pyramidal feature network [70] as the backbone. Features are then propagated and aggregated using GNN for target depth estimation. In [35], after obtaining the 2D poses from the 2D pose estimator, the poses were normalized to be centered on the root point. Then, the authors used three temporal models—joint-TCN, root-TCN, and velocity-TCN—to obtain absolute 3D human poses, but on monocular videos instead of single images.

On the other hand, bottom-up approaches [34,71,72] first produced all body joint locations and depth maps, then associated body parts to each person according to the root depth and part relative depth. Mehta et al. [34] proposed a single forward pass regardless of the number of people in the scene. The authors applied temporal and kinematic constraints in three steps to predict occlusion-robust PoseMaps (ORPM) and part affinity fields [27]. Another bottom-up multi-stage framework was proposed by Zanfir et al. [73], which first estimated the volumetric heatmaps to determine the 3D keypoint locations and limbs using the confidence scores of all possible connections, and then conducted skeleton grouping in order to assign limbs to various people. Likewise, Fabbri et al. [71] proposed estimating the volumetric heatmaps in an encoder–decoder manner. They first produced compressed volumetric heatmaps, which were used as ground truth, and then decompressed at test time to re-obtain the original representation. Zhen et al. [33] proposed estimating 2.5D representations of body parts first and then reconstructed the 3D human pose in a single-shot bottom-up framework. Wang et al. [74] also proposed distribution-aware single-stage models to represent 3D poses with a 2.5D human center, together with 3D center-relative joint offsets in a one pass solution.

TDBU_Net framework [36] combined top-down and bottom-up pipelines to accomplish the multi-person camera-centric 3D human pose estimation.

In this article, we were inspired by all of these proposals in building a top-down framework that could be used in real-world applications. We used monocular video as input, as in [35,36]. Thus, to deal with long-term models, we chose dilated temporal convolutional networks which only required the next images to produce real-time outputs. To respect this constraint, we also needed a system that integrated as few models as possible, unlike [35,36], while maintaining the highest possible accuracy.

3. Framework Overview

The first part of this section presents the basic architectures used in our framework, consisting of four phases: the human detector using Yolo-v3 architecture [75], the 2D human pose estimator employing HrNet network [23], the 3D root-relative pose estimator using the GAST-Net model [76], and the depth root estimator with the RootNet model [31]. The second part describes the overall pipeline of the framework. The last part details the series of enhancements of our framework on the 3D absolute pose estimator and their impacts on the final result.

3.1. Basic Models Architectures

Human detector (Yolo-v3): This architecture [75] predicts bounding boxes using dimension clusters as anchor boxes. The network predicts four coordinates for each bounding box (bbox): the 2D image coordinates of the top-left pixel of the bbox, the width and height of the bbox, and the confidence score. Darknet-53 was used for feature extraction.

2D pose estimator (HrNet): The high-resolution network [23] starts from a high-resolution subnetwork and gradually adds high-to-low resolution subnetworks one by one, by decreasing the resolution to half and increasing the width to double in separate branches that connect in parallel. In that way, high-resolution representation is maintained throughout the process. The input image size is 256 $\times$ 192 or 384 $\times$ 288, which produces 17 heatmaps (heatmap per each keypoint) of size 64 $\times$ 48 or 96 $\times$ 72 respectively. The authors proposed a small network (HRNet-W32) with 32 channels and a large one (HRNet-W48) with 48 channels.

3D root-relative pose estimator (GAST-Net): The majority of models that recently analyzed and interpreted input video were based on temporal convolutional networks (TCNs), which were initially introduced to action segmentation by Lea et al. [55]. The GAST-Net (graph attention spatiotemporal network) [76] is inspired by VideoPose3D [39]. The network predicts 3D poses from 2D keypoints. It is designed from dilated temporal convolutional networks (TCNs) to tackle long-term patterns and exhibit extended memory, and from a graph attention block that consists of two spatial attention networks. The local spatial attention network models the hierarchical and symmetrical structures of the human skeleton. The global spatial attention network adaptively extracts global semantic information to better encode the spatial characteristics of the human body.

Depth estimator (RootNet): Moon et al. [31] proposed a top-down system to estimate 3D multi-person poses from a single RGB image, consisting of human detection by the DetectNet model, absolute 3D human root localization by the RootNet model, and root-relative 3D single-person pose estimation by the PoseNet model. Both models adopt ResNet-50 pre-trained on the ImageNet dataset as a backbone to extract the global data. We are particularly interested in the RootNet model, which generates two outputs: the 2D image coordinates of the root's keypoint (x, y) estimated using soft-argmax on the root-heatmap (the central point of the individual), and the root depth absolute determined using a scalar value k, computed using focal lengths divided by the per-pixel distance factors and the human area ratio between the real-world and the image.

3.2. Taxonomy of the Framework

Given a sequence of bounding boxes from monocular RGB videos of a person or a group of people in real-time, the goal was to produce a sequence of 3D camera-centric coordinates of everyone in the scene. First, for each person, we assigned a unique ID i to be tracked through the successive frames. Then, we applied a high-resolution network (HRNet) [23] on each frame to produce 17 heatmaps. Each heatmap predicts 2D human joint locations in MS-COCO format P_{2D} for each detected individual.

The 2D-poses P^i_{2D} in 27 frames were collected and given thereafter to a 3D single-pose estimator, GAST-Net, for direct 2D-to-3D mapping and recovering of the 3D root-relative pose P^i_{3Drel}, where all produced joints were represented by their distances from the pelvis keypoints. GAST-Net was applied (as much as the number of people in the frame).

GAST-Net was chosen since it provides the best compromise between the number of frames required to process and the estimation precision. In fact, the methods with the best accuracies on monocular videos from Human3.6M (the largest database of 3D human pose estimation) are: temporal convolution [39] trained in semi-supervision learning, the Attention 3D Human Pose [77], which identifies significant frames and tensor outputs from each layer using the attention mechanism, the RIE paper [43], which improves the accuracy by relative information encoding that yields positional and temporal-enhanced representations, and Anatomy3D [78], which estimates the 3D skeleton by predicting bone orientation and length. These methods reached the MPJPEs (defined in Section 4) of 44.1, 43.3, 45.1, and 46.8 mm, respectively, but required 243 frames as input. This is very costly in terms of memory and processing time; moreover, this increases the delay between the image display and the result, which is not favorable for real-time processing. Furthermore, tracking several individuals on large time scales is more complicated and error-prone. On the other hand, approaches that employ few frames have higher errors. For example, VIBE [79] only used 16 frames but attained an MPJPE error of 65.6 mm, as well as TP-Net [80] which required 20 frames but had an average error of 52.1. Trajectory space factorization [41] scored an error of 46.6 mm from 50 frames; GAST-Net achieved an MPJPE of 46.2 mm using 27 frames. Thus, it presents a good compromise for use in real-world contexts.

For absolute depth estimation of the pelvis keypoint, we employed the RootNet network proposed in [31], due to its adaptability to any 3D root-relative estimator.

The proposed overall pipeline for estimating the absolute camera-centered coordinates of multi-person keypoints from a monocular camera is depicted in Figure 3. The pipeline comprises three boxes. Person detection and 2D keypoint estimation are included in the first box (green). The second box (orange) contains the 2D to 3D lift, and the last box (blue) contains the depth estimation.

3.3. 3D Absolute Pose Estimator

The purpose of this work was to develop a 3D multi-person camera-centric pose estimation system under industrial and real-world settings. Therefore, we started with a hybridization of well-chosen models, GAST-Net for predicting 3D root-relative keypoints and a RootNet network proposed in [31] for predicting absolute root depth (i.e., the depth of pelvis keypoint), obtained by multiplying k (defined above) by the scalar value of the network output. Then, the XY camera coordinates of the root were determined using the camera-intrinsic parameters, the image coordinates of the root, and the predicted absolute root depth. Finally, the absolute coordinates of the rest of the joints were estimated from these two predictions. We call this hybridization the GR method. On the MuPoTS-3D dataset, the system adopting the GR method outperformed previous methods by more than 12.1 percentage points on AP_{25}^{root}, contributing to more than 6.7 percentage points on 3D-PCK$_{abs}$. However, we observed that the root-relative keypoints were less good by 25.8 percentage points on PCK, which sparked the idea to upgrade the GAST-Net. While the original GAST-Net was trained on single-person databases [29], we chose to retrain our model on both a single-person video database (MPII-3DHP [81]) and a multi-person video database (MuCo-Temp [56]) with the required processing, following [56], to produce direct absolute keypoint coordinates. The TCN-based approaches evaluated on MuPoTS-3D were trained on the MPII-3DHP database, containing videos of a single person recorded in a green-screen studio and/or on the MuCo-3DHP database, composed of MPII-3DHP frames, containing multiple positions copied into a single frame. For this, in order to train the temporal networks, such as GAST-NET$_{ABS}$, [56] proposed MuCo-Temp, a temporal extension of MuCo-3DHP that was generated, such as MuCo-3DHP, but it is composed of videos instead of frames. As a result, the relative keypoint precision enhanced from 63.8% with the basic GAST-Net to 82.5% on PCK with our modified GAST-Net, which contributes to 1.6 percentage points in absolute points on 3D-PCK$_{abs}$ when compared to the first methodology of hybridization. Note that in the following we name the upgraded

GAST-Net by GAST-NET$_{ABS}$, and this methodology by the GA method. We noticed that although AP_{25}^{root} of GAST-NET$_{ABS}$ (measuring the root depth estimation) has improved compared to the state-of-the-art, it is still not as good as the first hybridization methodology . This pushed us to compute the root-relative keypoints from the absolute keypoints obtained by GAST-NET$_{ABS}$ and employ the RootNet for root depth estimation, generating final absolute joints. We call this methodology the GAR method. In this way, we increased the accuracy (compared to the literature approaches) by more than 8.8 percentage points on 3D-PCK$_{abs}$. Figure 4 presents the structural diagram of the various types of networks used in the framework.

All these experimental results will be presented, detailed, and analyzed in the next section (Section 4).

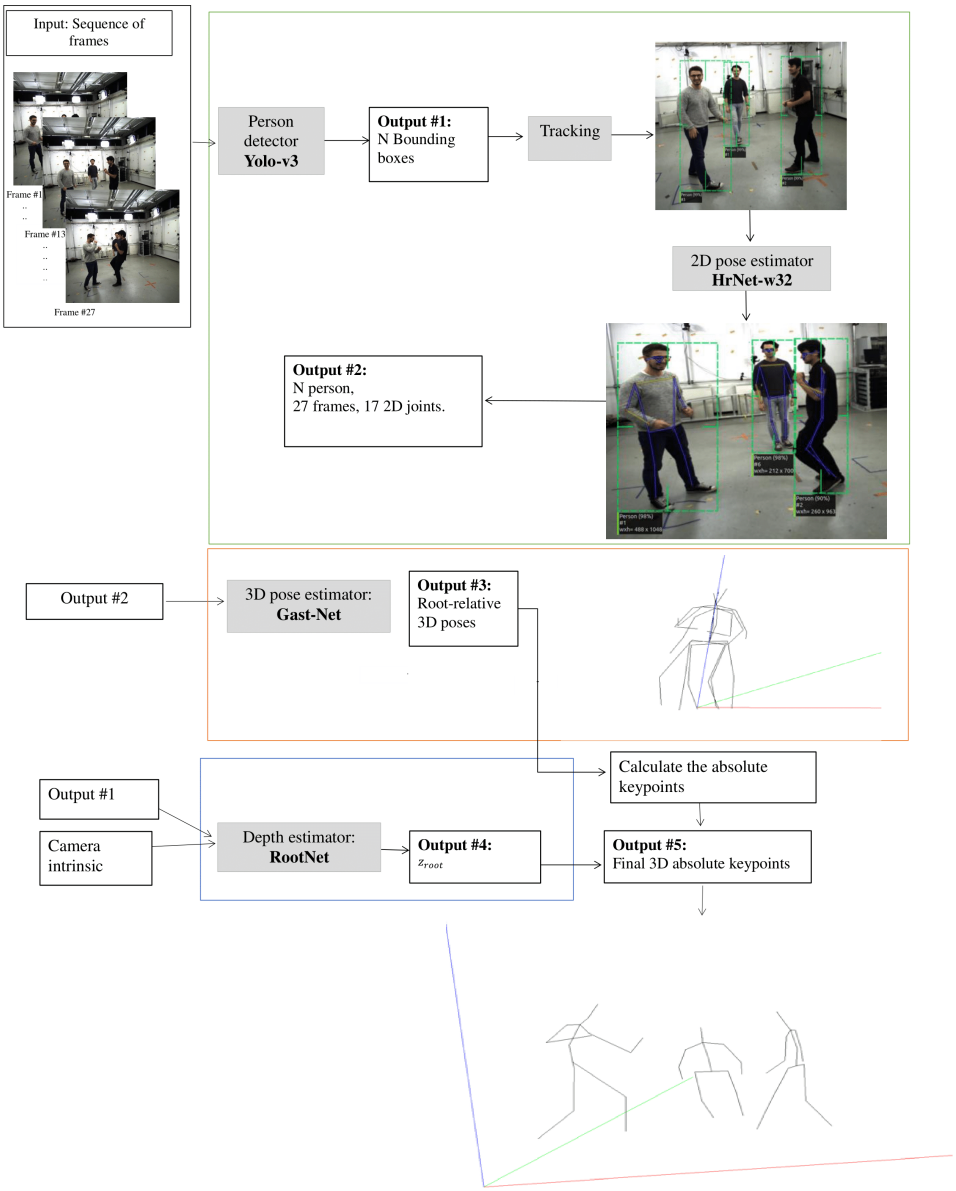

Figure 3. The pipeline of the Root-GAST-Net framework.

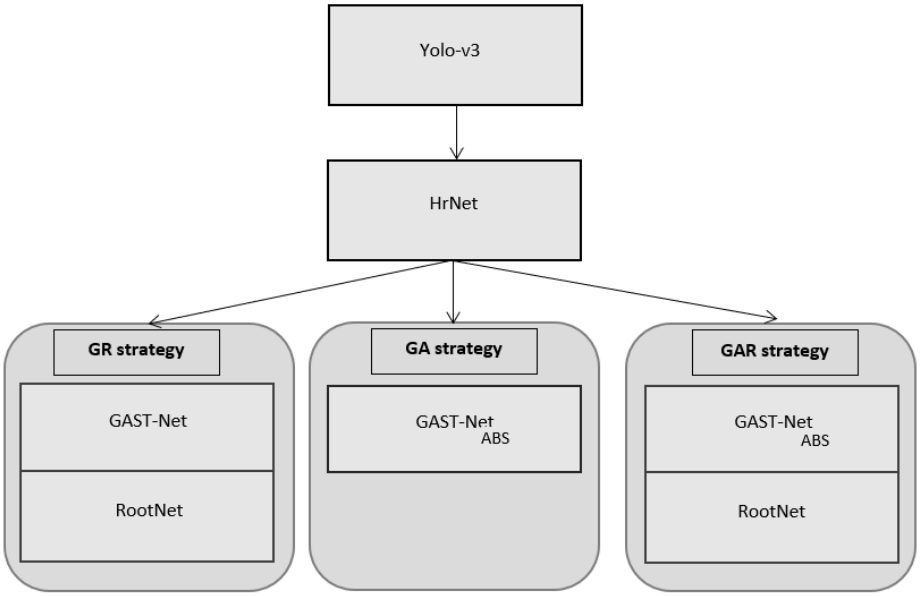

Figure 4. The structural diagram of the various types of networks used in the framework.

4. Experimentation and Results Discussion

This section deals with the experimental details and results of the proposed system. Results are discussed and evaluated using MPJPE, MRPE, 3D-PCK, AP_{25}^{root}, 3D-PCK_{abs} metrics and response times. The proposed Root-GAST-Net system and its three variants (GR, GA, GAR), 3D pose absolute methodologies, were compared to the existing methods grouped in papers_With_Code link of 3D multi-person pose estimations (absolutes) on the MuPoTS-3D page (https://paperswithcode.com/sota/3d-multi-person-pose-estimation-absolute-on, accessed on 1 April 2022). The compared methods are 3D MPPE PoseNet [31], HDNet [32], SMAP [33], HMOR [69], GnTCN [35], and TDBU_Net [36]. The goal of evaluating the three methodologies was to measure the impact of each adjustment.

4.1. Datasets and Evaluation Metrics

Human3.6M is the most popular and biggest dataset/benchmark for 3D human pose estimation [29]. It contains 3.6 million single-person indoor video frames and the corresponding poses of 11 professional actors (6 males, 5 females) captured by the MoCap system from 4 camera viewpoints. Camera extrinsic (rotation and translation with respect to world coordinates) and intrinsic parameters (focal length and principal point) are also available. This could be used to evaluate the single-person-centric pose estimate [39,41,43,57,77–80] as well as the camera-centered coordinate prediction [31–33,35,36,69]. Only subjects 9 and 11 were used for testing, as in prior studies.

For evaluation, we computed the mean per joint position error metric (MPJPE), which is the mean Euclidean error averaged over all joints and all poses, calculated after aligning the human root of the estimated and ground truth 3D poses, calculated on relative poses, as shown in the formula below:

$$MPJPE = \frac{1}{T}\frac{1}{N}\sum_{t=1}^{T}\sum_{i=1}^{N}\left\| J_i^{(t)} - J_i^{*(t)} \right\|_2, \quad (1)$$

where T denotes the total number of test samples and N denotes the number of joints. J and J^* denote the predicted joint and the ground truth joint, respectively.

Another evaluation metric used in this database, proposed in [31], is the mean root position error (MRPE), which is the average error of the absolute root joint (the hip) localization, as follows:

$$MRPE = \frac{1}{T}\sum_{t=1}^{T}\left\|\left(R^{(t)} - R^{*(t)}\right)\right\|_2, \quad (2)$$

where R and R^* denote the predicted root joint and the ground truth root joint respectively.

MuCo-3DHP and MuPoTS-3D MuCo-3DHP and MuPoTS-3D are two datasets proposed by Mehta et al. [34] for 3D multi-person pose estimation evaluation. MuCo-3DHP is the training dataset that merges randomly sampled 3D poses from a single-person 3D human pose dataset MPI-INF-3DHP [81] to form realistic multi-person scenes. MuPoTS-3D is a dataset used for testing 3D multi-person estimation. It contains 20 videos in both indoor and outdoor scenes. Ground truth is obtained with a multi-view markerless motion capture system.

In order to evaluate person-centric pose estimations, we used the percentage of a correct 3D keypoint (3D-PCK), which treats an estimated joint as correct if it is within a fixed threshold distance from the matched ground truth joint. In the literature, the threshold is set to 15 cm. We also used AUC_{rel}, which is the area under the 3D-PCK curve computed from various thresholds.

We followed [31] to evaluate the absolute camera-centered coordinate estimations. We used average precision AP_{25}^{root} to measure the 3D human root location prediction error, which considers the prediction as correct when the Euclidean distance between the estimated and the ground truth coordinates is smaller than 25 cm. Moreover, we used 3D-PCK_{abs}, which is PCK without the root alignment used to evaluate the absolute poses.

MuCo-Temp This dataset was proposed by [56]. It is generated in the same way as MuCo-3DHP. Both use images composited from the MPI-INF-3DHP dataset. The difference is that MuCo-Temp consists of videos instead of frames. So we can use it for temporal network training.

4.2. Implementation Details

We adopted Yolo-v3 architecture [75], which is based on the Darknet-53 model as a backbone and is pre-trained on the COCO dataset [82]. The input resolution is 608×608.

The cropped image of the bounding box was transformed to 384×288 to be used as input for the 2D pose estimator. The transformation applied was an affine transformation that preserves collinearity, parallelism, and the ratio of distances between the points, as in [23]. A unique ID was assigned to each person using the tracking method [83] based on the Hungarian optimization algorithm. Then, we used the small architecture of HrNet (HRNet-w32) pre-trained on the COCO dataset [82], implemented in PyTorch. The output was 17 heatmaps (resolution: 96×72). Cropping was resized to 256×256 to be processed by RootNet for depth root prediction Z_{abs}^{root}. A unique ID was affected for each person using the tracking method based on the Hungarian optimization algorithm. The 27 consecutive 2D coordinates were collected for each person, to be given to GAST-NET.

All networks, except GAST-NET, were optimized to TensorRT (https://developer.nvidia.com/tensorrt, accessed on 1 April 2022), a Nvidia library allowing to optimize computations on the GPU in order to reach lower computation times. This library also offers lower precision arithmetic but in our experiments, we kept models in the FP32 precision.

For GAST-NET_{ABS} training, we used the Adam optimizer with a learning rate of 1×10^{-3} and a batch size of 32. We trained the model for 80 epochs on MPII-3DHP [81] and MuCo-Temp [56] datasets. Computations were performed at the supercomputer facilities at Mésocentre Clermont Auvergne University for one week.

Finally, the detected bounding box was resized to 256×256 to be processed by RootNet for depth root prediction Z_{abs}^{root}.

4.3. Results

4.3.1. Evaluation of Multi-Person Dataset MuPoTS

The results of our system with the three improvements are listed in Table 1, which can be compared to the literature results. We evaluated using the MuPoTS-3D dataset since it

has been used to analyze 3D multi-person poses in both person-centric and camera-centric coordinates. Following [31,35], the performance of person-centric 3D pose estimation was evaluated using AUC_{rel} and PCK metrics, while camera-centric 3D pose estimation was evaluated using AP_{25}^{root} and PCK_{abs} metrics. The detailed PCK_{abs} results per sequence are shown in Table 2. We observed an improvement in the estimation accuracy in most of the sequences.

According to both tables, all our strategies outperformed previous 3D multi-person absolute pose estimation approaches by a significant margin, even if the relative poses were weaker.

Table 1. Person-centric and camera-centric evaluations on the MuPoTS-3D dataset. The best is in bold, the second best is underlined.

Method	**Year**	**PCK**	**AUC_{rel}**	**3D-PCK_{abs}**	**AP_{25}^{root}**
3D MPPE PoseNet [31]	2019	81.8	39.8	31.5	31.0
HDNet [32]	2020	83.7	-	35.2	39.4
SMAP [33]	2020	80.5	45.5	38.7	45.5
HMOR [69]	2020	82.0	43.5	43.8	-
GnTCN [35]	2021	<u>87.5</u>	<u>48.9</u>	45.7	45.2
TDBU_Net [36]	2021	**89.6**	**50.6**	48.0	46.3
DAS [74]	2022	82.7	-	39.2	-
Root-GAST with GR	-	63.8	30.6	54.7	<u>58.4</u>
Root-GAST with GA	-	82.5	45.3	<u>56.1</u>	56.8
Root-GAST with GAR	-	82.5	45.3	**56.8**	**58.9**

Table 2. Sequence-wise 3D-PCK_{abs} comparison with the state-of-the-art on the MuPoTS-3D dataset. (*) The accuracies of methods are measured on matched ground truths. The best is in bold, the second best is underlined.

Method	**S1**	**S2**	**S3**	**S4**	**S5**	**S6**	**S7**
3D MPPE PoseNet (*) [31]	59.5	45.3	51.4	46.2	53.0	27.4	23.7
HDNet [32]	21.4	22.7	58.3	27.5	37.3	12.2	49.2
SMAP (*) [33]	42.1	41.4	46.5	16.3	53.0	26.4	47.5
GnTCN (*) [35]	64.7	<u>59.3</u>	<u>59.4</u>	63.1	52.6	<u>42.7</u>	31.9
TDBU_Net [36]	<u>69.2</u>	57.1	49.3	<u>68.9</u>	<u>55.1</u>	36.1	<u>49.4</u>
Root-GAST with GAR (*)	**89.8**	**77.0**	**73.4**	**77.0**	**81.0**	**54.3**	**68.4**
Method	**S8**	**S9**	**S10**	**S11**	**S12**	**S13**	**S14**
3D MPPE PoseNet (*) [31]	26.4	39.1	23.6	8.3	14.9	38.2	29.5
HDNet [32]	<u>40.8</u>	<u>53.1</u>	43.9	43.2	**43.6**	39.7	28.3
SMAP (*) [33]	18.7	36.7	**73.5**	<u>46.0</u>	22.7	24.3	38.9
GnTCN (*) [35]	35.2	53.0	28.3	37.6	26.7	46.3	<u>44.5</u>
TDBU_Net [36]	33.0	43.5	52.8	**48.8**	<u>36.5</u>	<u>51.2</u>	37.1
Root-GAST with GAR (*)	**60.5**	**71.3**	<u>65.4</u>	33.5	26.1	**67.3**	**46.9**

Table 2. *Cont.*

Method	S15	S16	S17	S18	S19	S20	Avg
3D MPPE PoseNet (*) [31]	36.8	23.6	14.4	20.0	18.8	25.4	31.8
HDNet [32]	49.5	23.8	18.0	26.9	25.0	38.8	35.2
SMAP (*) [33]	47.5	34.2	35.0	20.0	38.7	64.8	38.7
GnTCN (*) [35]	50.2	47.9	39.4	23.5	**61.0**	**56.1**	46.3
TDBU_Net [36]	47.3	**52.0**	20.3	**43.7**	57.5	50.4	48.0
Root-GAST with GAR (*)	**66.9**	35.7	**40.1**	38.5	26.0	35.3	**56.8**

The average precisions throughout the entire dataset were then examined using various threshold settings ranging from 25 to 10 cm. AP measured the accuracy of the root key point; we only evaluated the Root-GAST system's performance using the GA approach since GR and GAR methodologies employed RootNet to predict the root joint. They produced the same result as the original paper. Table 3 displays the results. When compared to the state-of-the-art methodology, our method significantly achieves greater AP across all levels of thresholds. We deduce that our method estimates many more correct root keypoints even with a low distance threshold.

Table 3. Average precision of the root keypoint evaluation by different distances on the MuPoTS-3D dataset.

Method	AP_{25}^{root}	AP_{20}^{root}	AP_{15}^{root}	AP_{10}^{root}
3D MPPE PoseNet [31]	31.0	21.5	10.2	2.3
HDNet [32]	39.4	28.0	14.6	4.1
Root-GAST with GA	56.8	47.1	36.8	22.4

To compare with most of the existing methods that evaluate person-centric 3D pose estimations on MuPoTS-3D using MPJPE , we report our results using the same metric in Table 4. Our result was 101.9 mm, the result of [34] was 132 mm, the result of [84] was 120 mm, the result of [56] when adding the pose refinement model was 103 mm. Our method also outperforms the existing methods on this metric.

Table 4. MPJPE of the relative poses on the MuPoTS-3D dataset. The best is in bold, the second best is underlined.

Method	Year	MPJPE (mm)
Temporal smoothing [56]	2020	107
Temporal smoothing + Pose refinement [56]	2020	103
Depth Prediction Network [84]	2019	120
LCR-Net [67]	2017	146
Mehta et al. [34]	2018	132
GAST-Net$_{ABS}$	-	**101.9**

4.3.2. Evaluation on Single-Person Dataset Human3.6M

In order to validate the system, we chose Human3.6M, which contains only single-person videos. Since we compared the results through the mean root position error (MRPE) metric, which measured the accuracy error of the root key point, we only evaluated the Root-GAST system's performance using the GA approach. GR and GAR methodologies

employed RootNet to predict the root joint; they produced the same result as the original paper.

The root localization results of our GAST-Net$_{ABS}$ and the RootNet model are shown in Table 5. Even though the evaluation was performed on the Human3.6M dataset, we employed the GA model that was retrained on MPII and the MuCo-Temp dataset, and we compared it to the RootNet model that was trained on the MuCo dataset to make a fair comparison. Our measurement error amounted to 158 mm, while that of [31] was 289.28 mm. However, we could expect greater improvement if we train our model in the Human3.6M dataset.

Table 5. MRPE results comparison with RootNet [31] on the Human3.6M dataset. MRPE$_x$, MRPE$_y$, and MRPE$_z$ are the average MRPE errors in the x, y, and z axes, respectively.

Method	MRPE (mm)	MRPE$_x$ (mm)	MRPE$_y$ (mm)	MRPE$_z$ (mm)
3D MPPE PoseNet [31]	289.28	35.95	58.65	268.49
Root-GAST with GA	178	33	41.9	158

4.3.3. Response Time

The response time is the processing time taken by the algorithm to process its input; it depends on the material configurations. The Root-GAST-Net pipeline was implemented in C++ and executed on a machine equipped with Intel Core i5-9500, with a dedicated memory of 32GB, and the Nvidia GeForce GTX 1080, with a dedicated memory of 8GB.

A comparative analysis of the response times of each network is shown in Table 6. The processing time was measured on batches of monocular images from the Human3.6M dataset, each containing one person. Note that the processing time of the tracking step is negligible.

Table 6. Response time per model.

Model	Min Response Time (ms)	Max Response Time (ms)	Average Response Time (ms)
Yolo-v3	24	30	28
HrNet	9	12	10
GAST-Net	27	33	29
GAST-Net$_{ABS}$	23	29	26
RootNet	4	8	5

Finally, the frame rate of the whole pipeline with each strategy is given in Table 7. The proposed Root-GAST-Net system can run at about 15 frames per second, which is suitable for real-time scenarios. Therefore, improving the metrics does not impact the real-time aspect of the pipeline.

Table 7. Frame rate per strategy.

Strategy	Average Frame Rate (fps)
Root-GAST with GR	13
Root-GAST with GA	16
Root-GAST with GAR	15

4.3.4. Qualitative Results

As the system follows a top-down approach, the final result depends on all previous outputs. If the detection is not correctly done, the 2D keypoints and depths will be wrongly estimated, which will impact the absolute pose. If there are numerous people inside the box or body parts that are partially outside the box's bounds, the full-body joint calculation

is likely to be incorrect, as shown in Figure 5. The confusion stems from erroneous 2D point estimations, which have negative impacts on the 3D-lifted process.

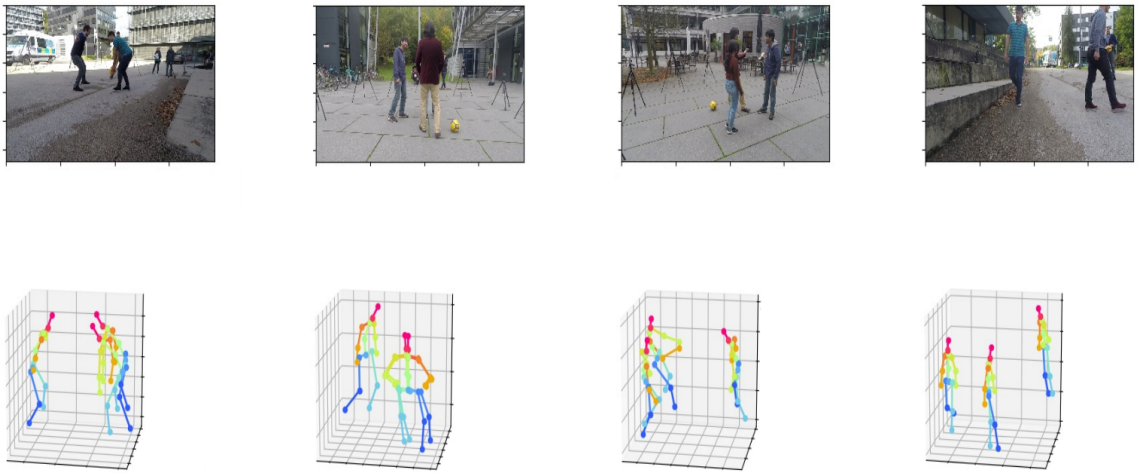

Figure 5. Erroneous 3D multi-person pose estimation. The first two images represent two similar poses of different people because one is completely occluded. In the right two images, one pose is incorrect because the body parts are partially outside of the boxes.

5. Conclusions

In this work, we propose a top-down framework for 3D multi-person absolute pose estimation, reconstructed from 2D poses from a monocular camera. Our framework Root-GAST-Net can combine different models in three strategies. The GR strategy and GAR strategy, which integrate human detection, 2D pose estimation, 3D human root-relative single-person pose estimation, and root depth estimation. Moreover, the GA strategy integrates human detection, 2D pose estimation, and 3D absolute pose estimation.

Experimental results on multiple datasets showed that our framework significantly outperforms the recent approaches in 3D absolute multi-pose estimation. In addition, the system can be used in real-time, as the execution time of each frame containing one person takes around 60 milliseconds using the Nvidia GeForce GTX 1080. This can be reduced using high-performance materials and FP16 precision.

In future works, we plan to retrain the model on the Human3.6M dataset to improve the evaluation accuracy of this database. We also plan to develop a fall detection application based on the absolute and relative 3D postures predicted by the Root-GAST-Net system.

Author Contributions: Conceptualization, all authors; methodology, all authors; software, A.E.K., D.B.; validation, all authors; formal analysis, all authors; investigation, A.E.K.; resources, all authors; data curation, A.E.K.; writing—original draft preparation, A.E.K.; writing—review and editing, D.B., V.B., K.B.; supervision, D.B., V.B., K.B.; project administration, D.B., V.B., K.B.; funding acquisition, V.B., K.B. All authors have read and agreed to the published version of the manuscript.

Funding: This work was funded by a CIFRE France/Morocco grant (2018/1635) with the Prynel Company, within a collaboration with University Mohammed V in Rabat, Morocco, financed by ANRT (Association Nationale de la Recherche et de la Technologie), France and CNRST (Centre National pour la Recherche Scientifique et Technique), Morocco.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: We used the MuCo-Temp dataset, generated with the GitHub project at pose_refinement (https://github.com/vegesm/pose_refinement (accessed on 12 February 2022)) for training. For evaluation, we used the Human 3.6M dataset, parsed and available at 3DMPPE_ROOTNET_RELEASE github project https://github.com/mks0601/3DMPPE_ROOTNET_RELEASE (accessed on

12 February 2022), and the MuPoTS-3D dataset is publicly available at website https://vcai.mpi-inf.mpg.de/projects/SingleShotMultiPerson/ (accessed on 12 February 2022).

Acknowledgments: The authors would like to thank Jérôme BROSSAIS for his fruitful contributions.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

HPE	human pose estimation
LSTM	long short-term memory
GNN	graph neural network
GCN	graph convolution network
TCN	temporal convolutional network
RNN	recurrent neural network
MPJPE	mean per joint position error
MRPE	mean of the root position error
AUC	area under the curve
3D-PCK	percentage of correct key-points in 3D space
AP^{root}	average precision of the root keypoint
GPU	graphics processing unit
GR	first 3D absolute pose methodology: GAST-Net + RootNet
GA	second 3D absolute pose methodology: GAST-Net_{ABS} trained on MuCo-Temp
GAR	third 3D absolute pose methodology: GAST-Net_{ABS} trained on MuCo-Temp + RootNet
Root-GAST	the whole pipeline: human detector + 2D pose estimator + 3D absolute pose estimator

References

1. Treleaven, P.; Wells, J. 3D body scanning and healthcare applications. *Computer* **2007**, *40*, 28–34. [CrossRef]
2. Grazioso, S.; Selvaggio, M.; Di Gironimo, G. Design and development of a novel body scanning system for healthcare applications. *Int. J. Interact. Des. Manuf.* **2018**, *12*, 611–620. [CrossRef]
3. Chromy, A.; Zalud, L. The RoScan thermal 3D body scanning system: medical applicability and benefits for unobtrusive sensing and objective diagnosis. *Sensors* **2020**, *20*, 6656. [CrossRef] [PubMed]
4. Liberadzki, P.; Adamczyk, M.; Witkowski, M.; Sitnik, R. Structured-light-based system for shape measurement of the human body in motion. *Sensors* **2018**, *18*, 2827. [CrossRef] [PubMed]
5. Nezami, F.N.; Wächter, M.A.; Maleki, N.; Spaniol, P.; Kühne, L.M.; Haas, A.; Pingel, J.M.; Tiemann, L.; Nienhaus, F.; Keller, L.; et al. Westdrive X LoopAR: An Open-Access Virtual Reality Project in Unity for Evaluating User Interaction Methods during Takeover Requests. *Sensors* **2021**, *21*, 1879. [CrossRef]
6. Ku Abd. Rahim, K.N.; Elamvazuthi, I.; Izhar, L.I.; Capi, G. Classification of human daily activities using ensemble methods based on smartphone inertial sensors. *Sensors* **2018**, *18*, 4132. [CrossRef]
7. Michonski, J.; Witkowski, M.; Sitnik, R.; Glinkowski, W.M. Automatic recognition of surface landmarks of anatomical structures of back and posture. *J. Biomed. Opt.* **2012**, *17*, 056015. [CrossRef]
8. Čibiraitė-Lukenskienė, D.; Ikamas, K.; Lisauskas, T.; Krozer, V.; Roskos, H.G.; Lisauskas, A. Passive detection and imaging of human body radiation using an uncooled field-effect transistor-based THz detector. *Sensors* **2020**, *20*, 4087. [CrossRef]
9. Reddy, N.D.; Guigues, L.; Pishchulin, L.; Eledath, J.; Narasimhan, S.G. TesseTrack: End-to-End Learnable Multi-Person Articulated 3D Pose Tracking. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 15190–15200.
10. He, Y.; Yan, R.; Fragkiadaki, K.; Yu, S.I. Epipolar transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 14–19 June 2020; pp. 7779–7788.
11. Iskakov, K.; Burkov, E.; Lempitsky, V.; Malkov, Y. Learnable triangulation of human pose. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27–28 October 2019; pp. 7718–7727.
12. Qiu, H.; Wang, C.; Wang, J.; Wang, N.; Zeng, W. Cross view fusion for 3d human pose estimation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27–28 October 2019; pp. 4342–4351.
13. Gordon, B.; Raab, S.; Azov, G.; Giryes, R.; Cohen-Or, D. FLEX: Parameter-free Multi-view 3D Human Motion Reconstruction. *arXiv* **2021**, arXiv:2105.01937.
14. Zhang, Y.; Wang, C.; Wang, X.; Liu, W.; Zeng, W. Voxeltrack: Multi-person 3d human pose estimation and tracking in the wild. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**. [CrossRef]

15. Tekin, B.; Márquez-Neila, P.; Salzmann, M.; Fua, P. Learning to fuse 2d and 3d image cues for monocular body pose estimation. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 3941–3950.
16. Moreno-Noguer, F. 3d human pose estimation from a single image via distance matrix regression. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2823–2832.
17. Lee, H.J.; Chen, Z. Determination of 3D human body postures from a single view. *Comput. Vision Graph. Image Process.* **1985**, *30*, 148–168. [CrossRef]
18. Zhou, X.; Zhu, M.; Pavlakos, G.; Leonardos, S.; Derpanis, K.G.; Daniilidis, K. Monocap: Monocular human motion capture using a cnn coupled with a geometric prior. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *41*, 901–914. [CrossRef] [PubMed]
19. Ghezelghieh, M.F.; Kasturi, R.; Sarkar, S. Learning camera viewpoint using CNN to improve 3D body pose estimation. In Proceedings of the 2016 Fourth International Conference on 3D Vision (3DV), Stanford, CA, USA, 25–28 October 2016; pp. 685–693.
20. Wu, J.; Xue, T.; Lim, J.J.; Tian, Y.; Tenenbaum, J.B.; Torralba, A.; Freeman, W.T. Single image 3d interpreter network. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; pp. 365–382.
21. Sun, X.; Xiao, B.; Wei, F.; Liang, S.; Wei, Y. Integral human pose regression. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 529–545.
22. Zhao, L.; Peng, X.; Tian, Y.; Kapadia, M.; Metaxas, D.N. Semantic graph convolutional networks for 3D human pose regression. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 3425–3435.
23. Sun, K.; Xiao, B.; Liu, D.; Wang, J. Deep high-resolution representation learning for human pose estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 5693–5703.
24. Li, J.; Su, W.; Wang, Z. Simple Pose: Rethinking and Improving a Bottom-up Approach for Multi-Person Pose Estimation. *arXiv* **2019**, arXiv:1911.10529.
25. Newell, A.; Yang, K.; Deng, J. Stacked hourglass networks for human pose estimation. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; pp. 483–499.
26. Chen, Y.; Wang, Z.; Peng, Y.; Zhang, Z.; Yu, G.; Sun, J. Cascaded pyramid network for multi-person pose estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7103–7112.
27. Cao, Z.; Simon, T.; Wei, S.E.; Sheikh, Y. Realtime multi-person 2d pose estimation using part affinity fields. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 7291–7299.
28. Zhang, F.; Zhu, X.; Dai, H.; Ye, M.; Zhu, C. Distribution-aware coordinate representation for human pose estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 14–19 June 2020; pp. 7093–7102.
29. Ionescu, C.; Papava, D.; Olaru, V.; Sminchisescu, C. Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Trans. Pattern Anal. Mach. Intell.* **2013**, *36*, 1325–1339. [CrossRef] [PubMed]
30. Sigal, L.; Balan, A.O.; Black, M.J. Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. *Int. J. Comput. Vis.* **2010**, *87*, 4. [CrossRef]
31. Moon, G.; Chang, J.Y.; Lee, K.M. Camera distance-aware top-down approach for 3d multi-person pose estimation from a single rgb image. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27–28 October 2019; pp. 10133–10142.
32. Lin, J.; Lee, G.H. Hdnet: Human depth estimation for multi-person camera-space localization. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; pp. 633–648.
33. Zhen, J.; Fang, Q.; Sun, J.; Liu, W.; Jiang, W.; Bao, H.; Zhou, X. Smap: Single-shot multi-person absolute 3d pose estimation. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; pp. 550–566.
34. Mehta, D.; Sotnychenko, O.; Mueller, F.; Xu, W.; Sridhar, S.; Pons-Moll, G.; Theobalt, C. Single-shot multi-person 3d pose estimation from monocular rgb. In Proceedings of the 2018 International Conference on 3D Vision (3DV), Verona, Italy, 5–8 September 2018; pp. 120–130.
35. Cheng, Y.; Wang, B.; Yang, B.; Tan, R.T. Graph and temporal convolutional networks for 3d multi-person pose estimation in monocular videos. *Proc. AAAI Conf. Artif. Intell.* **2021**, *4*, 12.
36. Cheng, Y.; Wang, B.; Yang, B.; Tan, R.T. Monocular 3D multi-person pose estimation by integrating top-down and bottom-up networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 7649–7659.
37. Cheng, Y.; Yang, B.; Wang, B.; Yan, W.; Tan, R.T. Occlusion-aware networks for 3d human pose estimation in video. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27–28 October 2019; pp. 723–732.
38. Cheng, Y.; Yang, B.; Wang, B.; Tan, R.T. 3d human pose estimation using spatio-temporal networks with explicit occlusion training. *Proc. AAAI Conf. Artif. Intell.* **2020**, *34*, 10631–10638. [CrossRef]
39. Pavllo, D.; Feichtenhofer, C.; Grangier, D.; Auli, M. 3d human pose estimation in video with temporal convolutions and semi-supervised training. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 7753–7762.

40. Chen, H.; Wang, Y.; Zheng, K.; Li, W.; Chang, C.T.; Harrison, A.P.; Xiao, J.; Hager, G.D.; Lu, L.; Liao, C.H.; et al. Anatomy-aware siamese network: Exploiting semantic asymmetry for accurate pelvic fracture detection in x-ray images. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; pp. 239–255.
41. Lin, J.; Lee, G.H. Trajectory space factorization for deep video-based 3d human pose estimation. *arXiv* **2019**, arXiv:1908.08289.
42. Li, W.; Zhao, Y.; Liu, Y.; Sun, M.; Waterhouse, G.I.; Huang, B.; Zhang, K.; Zhang, T.; Lu, S. Exploiting Ru-induced lattice strain in CoRu nanoalloys for robust bifunctional hydrogen production. *Angew. Chem.* **2021**, *133*, 3327–3335. [CrossRef]
43. Shan, W.; Lu, H.; Wang, S.; Zhang, X.; Gao, W. Improving Robustness and Accuracy via Relative Information Encoding in 3D Human Pose Estimation. In Proceedings of the 29th ACM International Conference on Multimedia, Nice, France, 21–25 October 2021; pp. 3446–3454.
44. Martinez, J.; Hossain, R.; Romero, J.; Little, J.J. A simple yet effective baseline for 3d human pose estimation. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2640–2649.
45. Fang, H.S.; Xu, Y.; Wang, W.; Liu, X.; Zhu, S.C. Learning pose grammar to encode human body configuration for 3d pose estimation. *Proc. AAAI Conf. Artif. Intell.* **2018**, *32*, 1.
46. Gong, K.; Zhang, J.; Feng, J. Poseaug: A differentiable pose augmentation framework for 3d human pose estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 8575–8584.
47. Zhou, X.; Zhu, M.; Leonardos, S.; Daniilidis, K. Sparse representation for 3D shape estimation: A convex relaxation approach. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *39*, 1648–1661. [CrossRef] [PubMed]
48. Zhou, X.; Zhu, M.; Leonardos, S.; Derpanis, K.G.; Daniilidis, K. Sparseness meets deepness: 3D human pose estimation from monocular video. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 4966–4975.
49. Chen, C.H.; Ramanan, D. 3d human pose estimation= 2d pose estimation+ matching. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 7035–7043.
50. Hossain, M.R.I.; Little, J.J. Exploiting temporal information for 3d human pose estimation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 68–84.
51. Lee, K.; Lee, I.; Lee, S. Propagating lstm: 3d pose estimation based on joint interdependency. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 119–135.
52. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef] [PubMed]
53. Zhang, H.; Shen, C.; Li, Y.; Cao, Y.; Liu, Y.; Yan, Y. Exploiting temporal consistency for real-time video depth estimation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27–28 October 2019; pp. 1725–1734.
54. Kumarapu, L.; Mukherjee, P. AnimePose: Multi-person 3D pose estimation and animation. *arXiv* **2020**, arXiv:2002.02792.
55. Lea, C.; Vidal, R.; Reiter, A.; Hager, G.D. Temporal convolutional networks: A unified approach to action segmentation. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; pp. 47–54.
56. Veges, M.; Lőrincz, A. Temporal Smoothing for 3D Human Pose Estimation and Localization for Occluded People. In Proceedings of the International Conference on Neural Information Processing, Bangkok, Thailand, 18–22 November 2020; pp. 557–568.
57. Liu, J.; Guang, Y.; Rojas, J. Gast-net: Graph attention spatio-temporal convolutional networks for 3d human pose estimation in video. *arXiv* **2020**, arXiv:2003.14179.
58. Mehta, D.; Sridhar, S.; Sotnychenko, O.; Rhodin, H.; Shafiei, M.; Seidel, H.P.; Xu, W.; Casas, D.; Theobalt, C. Vnect: Real-time 3d human pose estimation with a single rgb camera. *ACM Trans. Graph.* **2017**, *36*, 1–14. [CrossRef]
59. Cheema, N.; Hosseini, S.; Sprenger, J.; Herrmann, E.; Du, H.; Fischer, K.; Slusallek, P. Dilated temporal fully-convolutional network for semantic segmentation of motion capture data. *arXiv* **2018**, arXiv:1806.09174.
60. Li, W.; Liu, H.; Ding, R.; Liu, M.; Wang, P.; Yang, W. Exploiting temporal contexts with strided transformer for 3d human pose estimation. *IEEE Trans. Multimed.* **2022**. [CrossRef]
61. Yan, S.; Xiong, Y.; Lin, D. Spatial temporal graph convolutional networks for skeleton-based action recognition. In Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018.
62. Cai, Y.; Ge, L.; Liu, J.; Cai, J.; Cham, T.J.; Yuan, J.; Thalmann, N.M. Exploiting spatial-temporal relationships for 3d pose estimation via graph convolutional networks. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27–28 October 2019; pp. 2272–2281.
63. Qiu, Z.; Qiu, K.; Fu, J.; Fu, D. Dgcn: Dynamic graph convolutional network for efficient multi-person pose estimation. *Proc. AAAI Conf. Artif. Intell.* **2020**, *34*, 11924–11931. [CrossRef]
64. Zanfir, A.; Marinoiu, E.; Sminchisescu, C. Monocular 3d pose and shape estimation of multiple people in natural scenes-the importance of multiple scene constraints. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 2148–2157.
65. Rogez, G.; Weinzaepfel, P.; Schmid, C. Lcr-net++: Multi-person 2d and 3d pose detection in natural images. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *42*, 1146–1161. [CrossRef]
66. Pavlakos, G.; Zhou, X.; Derpanis, K.G.; Daniilidis, K. Coarse-to-fine volumetric prediction for single-image 3D human pose. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 7025–7034.

67. Rogez, G.; Weinzaepfel, P.; Schmid, C. Lcr-net: Localization-classification-regression for human pose. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 3433–3441.
68. Benzine, A.; Chabot, F.; Luvison, B.; Pham, Q.C.; Achard, C. Pandanet: Anchor-based single-shot multi-person 3d pose estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 14–19 June 2020; pp. 6856–6865.
69. Li, J.; Wang, C.; Liu, W.; Qian, C.; Lu, C. Hmor: Hierarchical multi-person ordinal relations for monocular multi-person 3d pose estimation. *arXiv* **2020**, arXiv:2008.00206.
70. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125.
71. Fabbri, M.; Lanzi, F.; Calderara, S.; Alletto, S.; Cucchiara, R. Compressed volumetric heatmaps for multi-person 3d pose estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7204–7213.
72. Zhang, C.; Zhan, F.; Chang, Y. Deep monocular 3d human pose estimation via cascaded dimension-lifting. *arXiv* **2021**, arXiv:2104.03520.
73. Zanfir, A.; Marinoiu, E.; Zanfir, M.; Popa, A.I.; Sminchisescu, C. Deep network for the integrated 3d sensing of multiple people in natural images. In Proceedings of the Advances in Neural Information Processing Systems 31 (NeurIPS 2018), Montréal, QC, Canada, 3–8 December 2018.
74. Wang, Z.; Nie, X.; Qu, X.; Chen, Y.; Liu, S. Distribution-Aware Single-Stage Models for Multi-Person 3D Pose Estimation. *arXiv* **2022**, arXiv:2203.07697.
75. Redmon, J.; Farhadi, A. Yolov3: An incremental improvement. *arXiv* **2018**, arXiv:1804.02767.
76. Liu, J.; Rojas, J.; Li, Y.; Liang, Z.; Guan, Y.; Xi, N.; Zhu, H. A graph attention spatio-temporal convolutional network for 3D human pose estimation in video. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 3374–3380.
77. Liu, R.; Shen, J.; Wang, H.; Chen, C.; Cheung, S.c.; Asari, V. Attention mechanism exploits temporal contexts: Real-time 3d human pose reconstruction. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 14–19 June 2020; pp. 5064–5073.
78. Chen, T.; Fang, C.; Shen, X.; Zhu, Y.; Chen, Z.; Luo, J. Anatomy-aware 3d human pose estimation with bone-based pose decomposition. *IEEE Trans. Circuits Syst. Video Technol.* **2021**, *32*, 198–209. [CrossRef]
79. Kocabas, M.; Athanasiou, N.; Black, M.J. Vibe: Video inference for human body pose and shape estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 14–19 June 2020; pp. 5253–5263.
80. Dabral, R.; Mundhada, A.; Kusupati, U.; Afaque, S.; Sharma, A.; Jain, A. Learning 3d human pose from structure and motion. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 668–683.
81. Mehta, D.; Rhodin, H.; Casas, D.; Fua, P.; Sotnychenko, O.; Xu, W.; Theobalt, C. Monocular 3d human pose estimation in the wild using improved cnn supervision. In Proceedings of the 2017 International Conference on 3D Vision, Qingdao, China, 10–12 October 2017; pp. 506–516.
82. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft coco: Common objects in context. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014; pp. 740–755.
83. Galčík, F.; Gargalík, R. Real-time depth map based people counting. In Proceedings of the International Conference on Advanced Concepts for Intelligent Vision Systems, Poznań, Poland, 28–31 October 2013; pp. 330–341.
84. Véges, M.; Lőrincz, A. Absolute human pose estimation with depth prediction network. In Proceedings of the 2019 International Joint Conference on Neural Networks, Budapest, Hungary, 14–19 July 2019; pp. 1–7.